ORIGINAL RESEARCH ARTICLE

Comparison of machine learning methods for the prediction of type 2 diabetes in primary care setting using EHR data

Amos Otieno Olwendo[1], *George Ochieng*[2], *Kenneth Rucha*[2]

[1]*Department: Health Records & Informatics, Jomo Kenyatta University of Agriculture & Technology, Nairobi, Kenya*

[2]*Department: Health Management & Informatics, Kenyatta University*

Corresponding Author, Email: *aolwend@jkuat.ac.ke*

ABSTRACT

Diabetes remains a major global public health challenge, thus the need for better methods for managing diabetes. Machine learning could provide reliable solutions to the need for early detection and management of diabetes. This study conducted experiments to compare a number of selected machine learning approaches to determine their suitability for early detection of diabetes in the primary care setting. A retrospective study was conducted using EHR dataset of confirmed cases of diabetes collected during routine care at Nairobi Hospital. Institutional ethical approvals were obtained, and data were retrieved from the database through stratified sampling based on gender. Diagnoses were confirmed using the ICD-10 codes. Records with 5% or so of missing values were excluded from this analysis. Data were processed by correction of errors and replacement of missing values using measures of central tendency. The data were transformed through normalization using the decimal-scaling method. Data analysis was conducted using selected supervised and unsupervised learning algorithms. Model performances were validated using metrics for the evaluation of classification and clustering results, respectively. Random Forest had the highest accuracy (0.95) and error rate (0.05), while Gradient Boosting and Multilayer Perceptron (MLP) with 3 hidden layers obtained accuracy (0.94) and error rate (0.06), respectively. The process of selecting machine learning algorithms needs to explore both supervised and unsupervised learning techniques. In addition, an appropriate architectural design of an MLP could present astounding results for classification tasks in primary care settings.

Keywords: Comparison, machine learning, classification, clustering, type 2 diabetes

1.0 Introduction

Diabetes mellitus is a major health challenge with an increasing prevalence worldwide. Diabetes mellitus is a group of metabolic conditions characterized by a defect in the secretion of insulin that results in either hyperglycemia or hypoglycemia, thus affecting the quality of health. Glucose is generated from the foods we eat in the bloodstream as a result of the actions of a hormone produced by the pancreas, insulin. Diabetes mellitus is classified as type 1 or type 2. Type 1 DM (T1DM), also known as insulin-dependent diabetes mellitus, is characterized by insulin deficiency and develops at any point during an individual's lifetime. The prevalence of T1DM is currently on the rise globally. Type 2 diabetes mellitus (T2DM), also referred to as non-insulin-dependent diabetes mellitus, affects millions worldwide and is characterized by a long-

ISSN 1561-7645 (online)
doi: 10.4314/jagst.v23i1.3

standing prediabetes state (Kumar, Kinyua and Kimotho, 2022). The other forms of diabetes include prediabetes, pancreoprive, and gestational diabetes. Prediabetes is a state in which the sugar level in the body is relatively high but not sufficiently high to be classified, as in the case of T2DM (Mehedi, Mollick, and Yasmin, 2022).

Diabetes inflicts a substantial financial and psychological burden on patients and members of their families. Presently, the diabetes condition cannot be reversed through treatment. However, the diabetes condition can be easily managed through early detection and management of the disease to prevent adverse disease effects. The development and progression of diabetes are characterized by a number of complications, which include cardiovascular diseases such as coronary heart disease with chest pain, heart attack, stroke, and atherosclerosis; neuropathy; nephropathy; retinopathy; skin infections; hearing impairment; Alzheimer's disease; preeclampsia; and macrosomia. Nutrition is an important factor influencing the risk of developing T2DM and, to some extent, T1DM. Excess availability of metabolites such as free fatty acids from (sources such as dark green leafy vegetables, olive oil, whole grain foods, and eggs) and branched-chain amino acids from (sources such as chicken, fish, eggs, beans, nuts, and soya) induces whole-body insulin resistance, thereby minimizing the development and progression of diabetes. In addition to diet, diabetes is best managed by maintaining an active lifestyle where one could be engaged, for example, in rigorous exercise or work. Diabetes is one of the diseases that require biomarker discovery and translation research to determine the clinical characteristics of their sub-phenotypes right from onset to the manifestation of its complications (Fritsch *et al.*, 2021; Olwendo, Ochieng, and Rucha, 2021).

Approximately 85% of deaths associated with diabetes experienced in middle- and low-income economies are a result of type 2 diabetes mellitus. In Sub-Saharan Africa, T2DM is becoming a problem due to socioeconomic progress that is resulting in changes in lifestyles since populations are adopting the consumption of processed foods due to busy schedules. The development of T2DM is preceded by a prediabetes state that is usually diagnosed late due to misdiagnosis. In Kenya, the prevalence of T2DM makes up approximately 92% of all diagnosed cases of diabetes. Studies conducted in the United States of America, China, and the United Kingdom, among others, show a common trend in the prevalence of prediabetes between 36 and 50%. As a result, there is a need for improving diagnostic and therapeutic strategies for the effective management of diabetes (Olwendo, Ochieng, and Rucha, 2020).

This study aims to appraise the performances of MLP with 1–5 hidden layers against selected supervised and unsupervised learning algorithms on a diabetes dataset. A number of studies have attempted to develop machine learning models for the early detection of diabetes. However, such studies have been biased in their attempts to compare the performances of MLP with other supervised learning methods. Multilayer perceptron, based on the configuration of the network, can produce admirable classification outcomes, yet a number of studies (De Silva, Jönsson, and Demmer, 2020; Yuk *et al.*, 2022) that have examined the performance of MLPs have hardly considered the fact that the performance of an MLP is dependent on the network architecture.

2.0 Materials and methods

2.1 Study design and sampling frame

This study implemented a retrospective cross-sectional study approved by the Nairobi Hospital Bioethics and Research Committee and conducted between May and December 2019. The study was also approved by the National Commission for Science, Technology, and Innovation (NACOSTI). The hospital databases contain digital records of patient information such as demographics, diagnoses, laboratory reports, and medication data. Data collection was conducted through the retrieval of data using structure query language (SQL) queries across the electronic health record (EHR) database of records of confirmed cases of diabetes mellitus. The diagnosis of diabetes was confirmed based on the International Classification of Disease (ICD) version 10 codes E10–E14. All the retrieved cases had been attended to during routine care between January 2012 and December 2016. Data were sampled through a stratified sampling technique, considering the gender balance in the dataset. The sampled records were for adult patients 18 years of age or older that did not contain any co-morbidities other than any of the categories of hypertension that were identified using ICD 10 codes I10–I15. Also, EHR data records with more than 5% missing values were not included in the sampled data.

2.2 Data processing

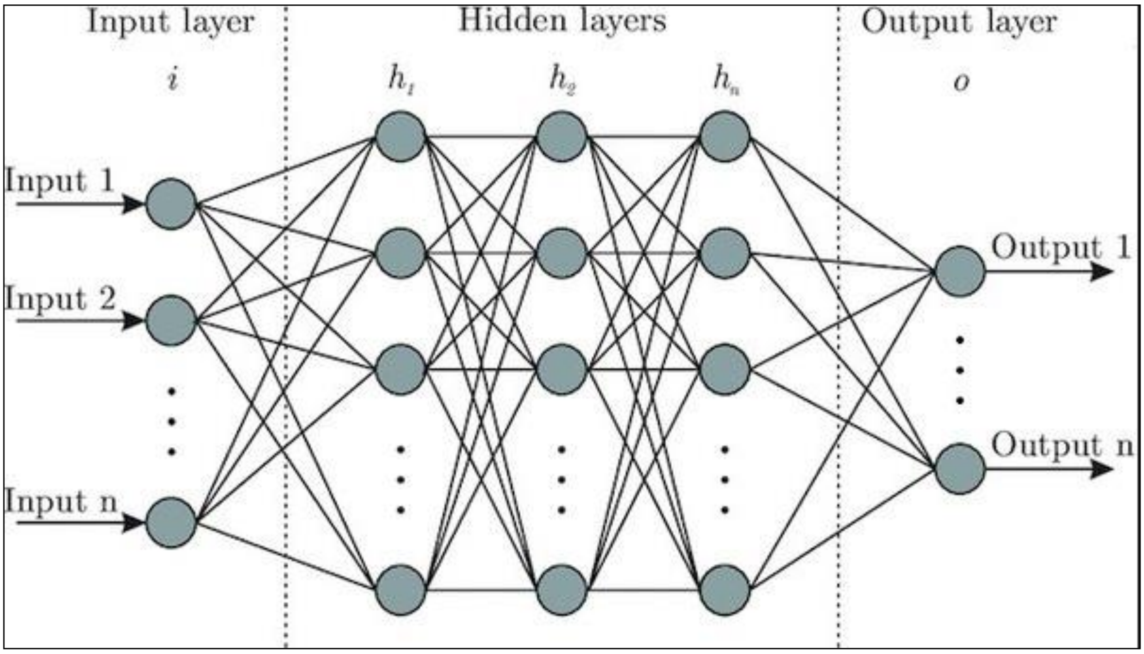
Data processing began with the de-identification of the records and the correction of data entry and typing errors using Knime Analytics software. Missing data values were replaced using measures of central tendency such as the mean and mode, respectively. Data were transformed through normalization using the decimal-scaling method and coded into a form appropriate for mining using machine learning algorithms. Thereafter, the random split method was applied, and the dataset was subdivided into equal proportions for the training and testing of the classification algorithms.

2.3 Experimental Setup

This experiment was conducted through the creation and configuration of 11 analytics models for the algorithms considered in this analysis. T2DM, the attribute for the confirmation of the presence or absence of Type 2 diabetes mellitus, was used as the predictor attribute for the supervised learning methods.

2.3.1 Supervised learning models

Data analysis was conducted by the development of selected supervised learning models: five multilayer perceptions (MLP) with 1 to 5 hidden layers and 10 hidden neurons per layer, probabilistic neural networks, gradient boosting, random forests, Naïve Bayes, support vector machines, and K-nearest neighbor classifiers. Artificial neural networks (ANN) are networks composed of simple elements operating in parallel. These elements are inspired by the biological nervous system, and the function of the network is determined by the connections between elements. Neural networks are trained to perform a particular task by adjusting the values of the connections between elements. MLP is a neural network architecture that is organized into layers: input, hidden, and output layers (Kihoro and Okango, 2014). Therefore, a multilayer feed-forward network performs a local adaptation of the weight updates according to the behavior of the error function. A probabilistic neural network (PNN) generates rules defined as high-dimensional Gaussian functions (Mehedi, Mollick, and Yasmin, 2022; Yuk *et al.*, 2022).



Source: (https://www.reddit.com/r/MachineLearning/comments/l1z8cr/d_best_way_to_draw_neural_network_diagrams/?rdt=34460)

Figure 1: A three hidden layer neural network architecture with n inputs/outputs.

2.3.1.1 Validation techniques for classification algorithms

Classification models are evaluated by measuring their sensitivity (the ability of the model to correctly identify positive cases) and specificity (the ability of the model to correctly identify negative cases). True positive (TP) means that the case is positive, and the model also found the given case to be positive. True negative (TN) means that a given case is negative, and the model also classified the same case as negative. False positive (FP) means that the case is negative and the model classified the case as positive. Finally, false negative (FN) means that the case is positive, but the classification model classified it as negative. Accuracy is a measure of the number of cases that have been correctly classified, divided by the total number of test cases. Accuracy is calculated using the formulae; Accuracy = (TP + TN) / (FP + TN + FP + FN).

Precision is a metric used to identify the correctness of the classification results. Precision is calculated as Precision = TP/(TP + FP).

Recall helps us determine the number of positive cases that are correctly identified out of the total number of positive cases. A recall is calculated as TP/(TP + FN).

The F1-score is a harmonic mean of recall and precision. The F1 score is calculated as 2 * ((Precision * Recall) / (Precision + Recall)).

Finally, another metric for the performance of a classification algorithm is to determine the proportion of misclassification of the cases in the dataset, also known as the error rate. The error rate is calculated as (FP + FN) / (TP + TN + FP + FN) (Shahmoradi *et al.*, 2017).

2.3.2 Unsupervised learning models

Clustering is an unsupervised machine learning method for partitioning a dataset into groups known as clusters. The clustering algorithms examined in this study, including models of fuzzy C-means, K-means, and DBSCAN algorithms, were further cross-examined. Fuzzy C-means is a clustering algorithm that can be used to reveal the underlying structure of the data. Fuzzy C-means allows data points to belong to more than one cluster, with a degree of membership in each of the clusters it belongs to. Also, K-means is a clustering algorithm that performs crisp clustering that assigns a data vector to exactly one cluster. The task of clustering stops when the cluster assignments for the data points do not change any longer. Finally, DBSCAN is a density-based clustering algorithm that defines three types of points in a dataset. The core points are points that have at least a minimum number of neighbors (MinPts) within a specified distance (eps). Border points are points that are within eps or a core point but have fewer than MinPts neighbors. Noise points are neither core points nor border points. DBSCAN builds clusters by joining the core points to one another. If a core point is within eps of another core point, they are termed directly density-reachable. All points that are within eps of a core point are termed density-reachable and are considered to be part of a cluster. The rest of the data points that are considered not density-reachable are all considered to be noise (Pekel and Özcan, 2018; Olwendo, Ochieng, and Rucha, 2021; Mehedi, Mollick, and Yasmin, 2022).

2.3.2.1 Validation techniques for clustering algorithms

The performance of unsupervised learning models is validated using two methods: internal validation, which looks into the cohesion (compactness and connectedness) of the clustered data points within a given cluster, and separation between different clusters. Cohesion for a cluster is computed by summating the similarity between each pair of records contained in the cluster.

$$\text{Cohesion } (C_k) = \underset{x \text{ € } C_k \,;\, y \text{ € } C_k}{\text{Similarity } (x,y)}$$

The separation between two clusters can be computed by summating the distance between each pair of records falling within the two clusters, and both records are from different clusters.

$$\text{Separation } (Cj,Ck) = \sum_{x\text{€}C_j;\ y\text{€}C_k} \text{Similarity } (x,y)$$

A set of clusters having high cohesion within the clusters and high separation between the clusters is considered to be good, or rather well-formed. In addition, silhouette analysis measures the extent to which a given data point is clustered. Silhouette analysis estimates the average distance between clusters. Silhouette coefficients (S_i) range between -1 and 1. The silhouette algorithms can be summarised as:

For a given observation i, Silhouette width s_i is calculated as;

i. For a given observation *i*, calculate the average dissimilarity between *i* and all other points of the cluster to which *i*

ii. For all other clusters to which *i* do not belong, calculate the average dissimilarity for *i* in all the observations of C.

iii. Then calculate the silhouette width of the given observation.

Therefore, coefficients with S_i close to 1 are considered to be well clustered; S_i close to 0 means that the given data point lies between two clusters; and S_i with a negative value means that these data points have been placed in the wrong cluster.

3.0 Results

3.1 Description of the EHR dataset

The dataset comprised 652 records of diabetes mellitus. Records of the female gender were 372/652 (57%), and the average age and BMI of the patients were 53 years and 30 years, respectively. Also, the averages of the systolic and diastolic blood pressures were 133 and 81 mm Hg, respectively. Furthermore, the dataset was comprised of 92% confirmed cases of T2DM. The other details are summarised in Table 1.

Table 1: Description of the EHR dataset for cases diabetes mellitus

Attribute	Attribute Description	Range and Value
Age	Age of the patient (in years)	21 to 82
Gender	Gender of the patient	0,1 (0: Female, 1: Male)
BMI	Body mass index of the patient	17 to 45
Pulse	Measured pulse rate of the patient	55 to 118
Systolic	Measured systolic blood pressure of the patient	70 to 203
Diastolic	Measured diastolic blood pressure of the patient	45 to 111
RandomBS	Measured random blood sugar of patient	2 to 22
SPO2	Measured saturation of oxygen in the blood of the patient	10 to 100
Temperature	Patient's measured body temperature	20 to 37
Respiration	Patient's respiration rate	12 to 22
HTN	Diagnosis for hypertension based on ICD 10 code	0,1 (0: Absent, 1: Present)
T1DM	Diagnosis for Type 1 diabetes mellitus based on ICD 10 code	0,1 (0: Absent, 1: Present)
T2DM	Diagnosis for Type 2 diabetes mellitus based on ICD 10 code	0,1 (0: Absent, 1: Present)

3.1.1 Variable significance analysis on the state of T2DM

The analysis of the significance of the independent variables (features) on the state of the dependent variable (state of T2DM) was determined by the calculation of the area under the curve for the state variable T2DM set to a value of 1 (present), which shows that variables such as body temperature (0.460), presence or absence of T1DM (0.462), and respiration (0.456) are less significant in the determination of the outcome (presence or absence of T2DM). Moreover, features such as age, pulse rate, and the presence of hypertension are essential factors in the diagnosis of type 2 diabetes mellitus. The rest of the details are summarised in Table 2.

Table 2: Area Under the Curve (ROC) with the state variable (T2DM) set to 1.

Variable	Area
Age	.606
Body mass index	.521
Pulse rate	.626
Systolic blood pressure	.571
Diastolic blood pressure	.558
Random blood sugar	.571
SPO2	.508
Hypertension	.600
Gender	.568
Temperature	.460
Type 1 diabetes mellitus	.462
Respiration	.456

3.2 Sensitivity and specificity of the eight supervised learning models

The dataset was split into two equal halves of 326 for the training dataset and the testing dataset, respectively. A multilayer perceptron with 1 and 2 (MLP1 and MLP2) hidden layers reported the highest sensitivity; 301/326 (92.3%) of the test dataset had been correctly identified as true cases of T2DM. Also, MLP1 and MLP2 models reported the highest proportions of type II error (false negative): 25/326 (0.08%). On the other hand, the Naïve Bayes algorithm reported the highest type I error at 213/326 (65%), followed by the support vector machine at 26/326 (0.08%). The other details are summarised in Table 3.

Table 3: A summary of the sensitivity and specificity of the eight supervised models

Model	No. of Hidden Layers	TP	FP	TN	FN
Multilayer Perceptron	1	301	0	0	25
	2	301	0	0	25
	3	299	2	9	16
	4	295	7	8	16
	5	290	11	9	16
Probabilistic Neural Network		300	1	1	23
Gradient Boosting		299	1	9	17
Random Forest		297	12	14	3
Naïve Bayes		25	213	87	1
Support Vector Machine		300	26	0	0
K-Nearest Neighbour		298	2	5	21

3.3 Model performance for the supervised learning algorithms

The random forest model reported the highest accuracy (0.95) and the lowest error rate (0.05), while the Naïve Bayes algorithm reported the lowest accuracy (0.34) and the highest error rate (0.66). Also, the multilayer perceptron with 3 hidden layers reported similar results to the gradient boosting model with accuracy (0.94) and an error rate (0.06). The other details are summarised in Table 4.

Table 4: Model performance for the supervised learning algorithms

Model	No. of Hidden Layers	Accuracy	Precision	Recall	F1-Score	Error rate
Multilayer Perceptron	1	0.92	1.00	0.92	0.96	0.08
	2	0.92	1.00	0.92	0.96	0.08
	3	0.94	0.99	0.95	0.97	0.06
	4	0.93	0.98	0.95	0.96	0.07
	5	0.92	0.96	0.95	0.96	0.08
Probabilistic Neural Network		0.92	0.99	0.93	0.96	0.07
Gradient Boosting		0.94	1.00	0.95	0.97	0.06
Random Forest		0.95	0.96	0.99	0.98	0.05
Naïve Bayes		0.34	0.11	0.96	0.19	0.66
Support Vector Machine		0.92	0.92	1.00	0.96	0.08
K-Nearest Neighbor		0.93	0.99	0.93	0.96	0.07

3.4 Model performance for the K-Means algorithm

The performance of the k-means model in the clustering task for 652 records of cases of diabetes into four clusters numbered cluster_0 to cluster_3 was optimal, with 99.7% assigned a positive silhouette coefficient value. The number of records wrongfully clustered (with a negative silhouette coefficient) was 0.03%. The other details are summarised in Table 5 and Figure 1, respectively.

Table 5: Evaluation of the performance of the k-means model based Silhouette coefficients

Cluster N= 652	Positive Silhouette Coefficient	Negative Silhouette Coefficient
cluster_0	23.3%	0.00%
cluster_1	25.0%	0.00%
cluster_2	25.8%	0.00%
cluster_3	25.6%	0.30%

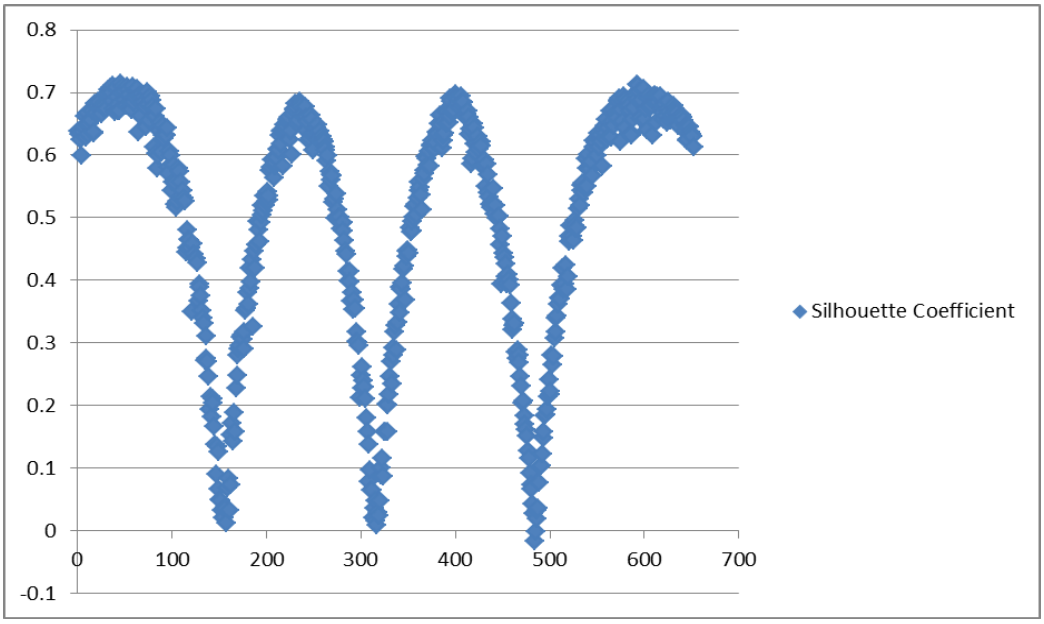


Figure 2: A scatter plot of the values of Silhouette coefficients by the k-means model

3.5 Model performance for the Fuzzy C-Means algorithm

The k-means model correctly clustered 71.9% of the diabetes dataset. Also, 28.1% of the records that had been clustered as noise had a negative silhouette coefficient. This variation in the silhouette coefficient values for all 652 records of cases of diabetes is observable, as summarised in Figure 2.

Table 6: Evaluation of the performance of the Fuzzy c-means model based on Silhouette coefficients

Cluster	Positive Silhouette Coefficient N = 652	Negative Silhouette Coefficient
Noise	0.0%	28.1%
cluster_0	23.9%	0.0%
cluster_1	23.6%	0.0%
cluster_2	24.4%	0.0%

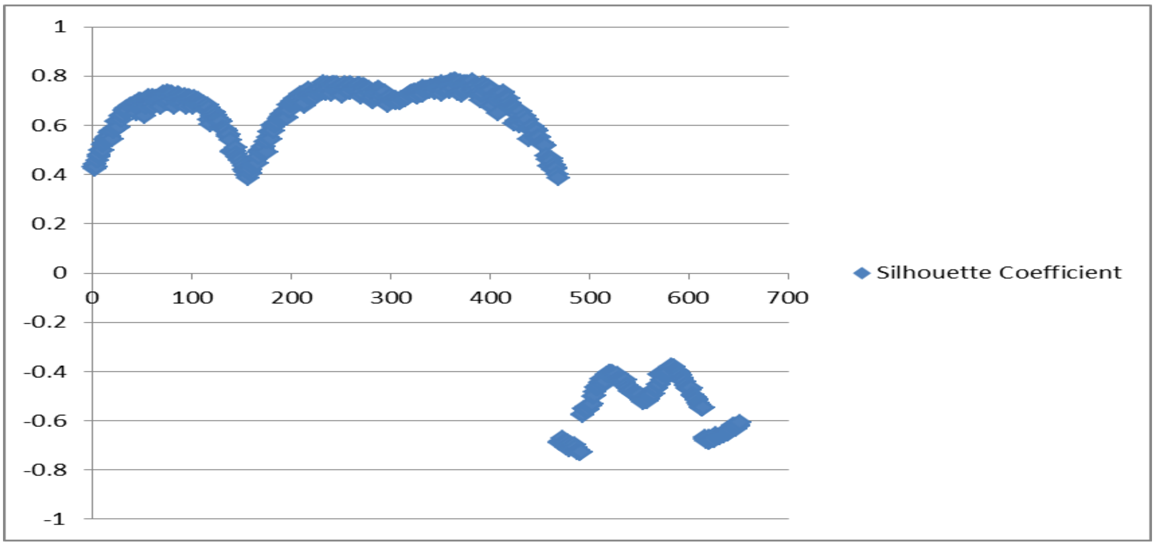


Figure 3: A scatter plot of the values of Silhouette coefficients by the Fuzzy c-means model

4.0 Discussion

This study appraised the performances of a selected set of supervised learning models: multilayer perceptron with 1–5 hidden layers, probabilistic neural network, random forest, gradient boosting, Naïve Bayes, support vector machine, and k-nearest neighbour algorithms on a diabetes dataset. Furthermore, the same dataset was also subjected to the three selected unsupervised learning algorithms: DBSCAN, k-means, and fuzzy c-means. Results from the clustering models were also validated by the calculation of the cohesion within each cluster and the separations between the different clusters by determining the silhouette coefficients for every case.

Variables significant in determining the state of Type 2 diabetes mellitus include body mass index, pulse rate, and systolic and diastolic blood pressure. However, this study mainly focused on demographic characteristics or non-laboratory (Dong, Cheng, *et al.*, 2022; Dong, Tse, *et al.*, 2022) features that are easily accessible whenever a person presents themselves at the triage for examination. A number of studies (Yuk *et al.*, 2022) tend to consider features that require laboratory tests. However, such an approach could be very inconvenient in emergencies and also in rural settings (Birk *et al.*, 2021), where laboratory services may be hard to come by.

Artificial neural networks remain one of the top-performing classification algorithms, and the results produced definitely depend on their architectural design (Fitria, Yulisda, and Ula, 2021). Even though the overall performance of the random forest seemed to be better than all the MLP models, the MLP with 3 hidden layers had a lower type I error compared to the case of the random forest. However, the random forest algorithm presented the lowest type II error (Abdollahi and Nouri-Moghaddam, 2022). Thus, the selection of machine learning should be adequately informed not only by the accuracy of the model but also by the model's performance on type I and II errors, in addition to other metrics such as accuracy and the F1 score.

Furthermore, the process of selecting a machine learning algorithm should also be guided by the nature of the task at hand. In this study, the popular Naïve Bayes classifier performed so poorly with an error rate of 0.66 (Fitria, Yulisda, and Ula, 2021; Abdollahi and Nouri-Moghaddam, 2022). On the other hand, the results of the DBSCAN algorithm were also very shocking, as all 652 records of cases of diabetes were classified as noise. According to the DBSCAN algorithm, which is popularly known for working with noisy data, there were no meaningful patterns of clusters in the dataset. Could such ambiguity in the dataset be the same reason for the poor performance of the Naïve Bayes classifier?

Nonetheless, the K-means and fuzzy c-means algorithms identified meaningful clusters from the dataset. As a matter of fact, Silhouette's analysis of the clusters of both the k-means and the fuzzy c-means reported that only about 30% had negative coefficients, meaning that a larger percentage of the dataset had been placed in the clusters they relatively belonged to. Therefore, a comparative analysis of the performances of machine learning methods needs to consider sampling from both the supervised and unsupervised algorithms and compare results within each of the two categories (Olwendo, Ochieng, and Rucha, 2020; Mehedi, Mollick, and Yasmin, 2022).

This study successfully conducted an experiment to compare the performance of selected supervised learning models. The random forest, gradient-boosted tree, and MLP with 3 hidden layers presented the best results. However, this study had a number of limitations in comparison to a number of studies relevant to this topic. The first limitation was a small sample size of 652, while the other studies that conducted similar analyses, such as Dong, Tse, *et al.* (2022; Hu, Lai, and Farid (2022; Zueger *et al.* (2022) used sample sizes between 1800 and 22000. Larger sample sizes should not substantially affect results. However, the larger the sample size, the better the training for supervised learning algorithms; thus, the distinctions in the performances of supervised algorithms will be more obvious. The second limitation was that this study was conducted using historical routine healthcare data collected between January 2012 and December 2016, and we had no knowledge of who collected the data or the criteria used for the identification of cases of diabetes. However, given the fact that relevant ICD-10 codes were used to confirm the diagnosis of diabetes and hypertension, it is generalizable that standard diagnostic protocols were followed.

5.0 Acknowledgements

5.1 Funding

None

5.2 General acknowledgement

None

5.3 Conflict of interest

None.

7.0 References

Abdollahi, J. and Nouri-Moghaddam, B. (2022) 'Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction', *Iran Journal of Computer Science*, 5(3), pp. 205–220. doi.org/10.1007/s42044-022-00100-1

Birk, N. *et al.* (2021) 'Exploration of Machine Learning and Statistical Techniques in Development of a Low-Cost Screening Method Featuring the Global Diet Quality Score for Detecting Prediabetes in Rural India', *The Journal of Nutrition*, 151(Supplement_2), pp. 110S-118S. doi.org/10.1093/jn/nxab281

De Silva, K., Jönsson, D. and Demmer, R.T. (2020) 'A combined strategy of feature selection and machine learning to identify predictors of prediabetes', *Journal of the American Medical Informatics Association*, 27(3), pp. 396–406. doi.org/10.1093/jamia/ocz204

Dong, W., Cheng, W.H.G., *et al.* (2022) 'Development and validation of a diabetes mellitus and prediabetes risk prediction function for case finding in primary care in Hong Kong: a cross-sectional study and a prospective study protocol paper', *BMJ Open*, 12(5), p. e059430. doi.org/10.1136/bmjopen-2021-059430.

Dong, W., Tse, T.Y.E., *et al.* (2022) 'Non-laboratory-based risk assessment model for case detection of diabetes mellitus and pre-diabetes in primary care', *Journal of Diabetes Investigation*, 13(8), pp. 1374–1386. doi.org/10.1111/jdi.13790

Fitria, R., Yulisda, D. and Ula, M. (2021) 'DATA MINING CLASSIFICATION ALGORITHMS FOR DIABETES DATASET USING WEKA TOOL', *Sisfo: Jurnal Ilmiah Sistem Informasi*, 5(2). doi.org/10.29103/sisfo.v5i2.6236

Fritsche, A. *et al.* (2021) 'Different Effects of Lifestyle Intervention in High- and Low-Risk Prediabetes: Results of the Randomized Controlled Prediabetes Lifestyle Intervention Study (PLIS)', *Diabetes*, 70(12), pp. 2785–2795. Available at: https://doi.org/10.2337/db21-0526

Hu, H., Lai, T. and Farid, F. (2022) 'Feasibility Study of Constructing a Screening Tool for Adolescent Diabetes Detection Applying Machine Learning Methods', *Sensors*, 22(16), p. 6155. Available at: https://doi.org/10.3390/s22166155

Kihoro, J.M. and Okango, E.L. (2014) 'STOCK MARKET PRICE PREDICTION USING ARTIFICIAL NEURAL NETWORK: AN APPLICATION TO THE KENYAN EQUITY BANK SHARE PRICES', *JOURNAL OF AGRICULTURE, SCIENCE AND TECHNOLOGY*, 16(1), pp. 161–172.Available at: https://ojs.jkuat.ac.ke/index.php/JAGST/article/view/76

Kumar, B., Kinyua, J. and Kimotho, J. (2022) 'Determination of Microbial Metagenomic Markers of Type 2 Diabetes Mellitus (T2DM) in Patients Visiting South C Health Centre in Nairobi Kenya', JOURNAL OF AGRICULTURE, SCIENCE AND TECHNOLOGY, 21(2), pp. 83–95. Available at: https://doi.org/10.4314/jagst.v21i2.7

Mehedi, H., Mollick, S. and Yasmin, F. (2022) 'An unsupervised cluster-based feature grouping model for early diabetes detection', *Healthcare Analytics*, 2, p. 100112. Available at: https://doi.org/10.1016/j.health.2022.100112

Olwendo, A.O., Ochieng, G. and Rucha, K. (2020) 'Prevalence and Complications Associated with Diabetes Mellitus at the Nairobi Hospital, Nairobi City County, Kenya', *Journal of Health Informatics in Africa*, 7(2), pp. 47–57. Available at: https://doi.org/10.12856/JHIA-2020-v7-i2-290

Olwendo, A.O., Ochieng, G. and Rucha, K. (2021) 'Suitability of Electronic Health Record Data for Computational Phenotyping of Diabetes Mellitus at Nairobi Hospital, Nairobi City County, Kenya | East African Journal of Science, Technology and Innovation', 2(2). Available at: https://doi.org/10.37425/eajsti.v2i2.224.

Pekel, E. and Özcan, T. (2018) 'DIAGNOSIS OF DIABETES MELLITUS USING STATISTICAL METHODS AND MACHINE LEARNING ALGORITHMS', *Sigma Journal of Engineering and Natural Sciences*, 36(4), pp. 1265–1282. https://dergipark.org.tr/en/pub/sigma/issue/65501/1013518

Shahmoradi, L. *et al.* (2017) 'A Probabilistic Model for COPD Diagnosis and Phenotyping Using Bayesian Networks', *Journal of Community Health Research*, 6(1), pp. 34–43.

Yuk, H. *et al.* (2022) 'Artificial Intelligence-based Prediction of Diabetes and Prediabetes Using Health Checkup Data in Korea', 36(1), p. 3772. https://www.tandfonline.com/doi/full/10.1080/08839514.2022.2145644

Zueger, T. *et al.* (2022) 'Machine Learning for Predicting the Risk of Transition from Prediabetes to Diabetes', *Diabetes Technology & Therapeutics*, 24(11), pp. 842–847. Available at: https://doi.org/10.1089/dia.2022.0210